Motivational Factors in Computer-administered Integrated Skills Tasks

Kormos, Judit; Brunfaut, Tineke; Michel, Marije

Published in:
Language Assessment Quarterly

DOI:
10.1080/15434303.2019.1664551

Routledge
Taylor & Francis Group

Language Assessment Quarterly

ISSN: 1543-4303 (Print) 1543-4311 (Online) Journal homepage: https://www.tandfonline.com/loi/hlaq20

Motivational Factors in Computer-administered Integrated Skills Tasks: A Study of Young Learners

Judit Kormos, Tineke Brunfaut & Marije Michel

To cite this article: Judit Kormos, Tineke Brunfaut & Marije Michel (2020) Motivational Factors in Computer-administered Integrated Skills Tasks: A Study of Young Learners, Language Assessment Quarterly, 17:1, 43-59, DOI: 10.1080/15434303.2019.1664551

To link to this article: https://doi.org/10.1080/15434303.2019.1664551

LANGUAGE ASSESSMENT QUARTERLY
2020, VOL. 17, NO. 1, 43–59
https://doi.org/10.1080/15434303.2019.1664551

Routledge
Taylor & Francis Group

Motivational Factors in Computer-administered Integrated Skills Tasks: A Study of Young Learners

Judit Kormos[a], Tineke Brunfaut[a], and Marije Michel[a,b]

[a]Lancaster University, Lancaster, United Kingdom of Great Britain and Northern Ireland; [b]Rijksuniversiteit Groningen, Groningen, Netherlands

ABSTRACT

Previous studies examined the association between motivational characteristics and language learning achievement, but considerably less is known about young language learners' task-specific motivation in assessment contexts. Our study investigated the task motivation of young learners of English when completing computer-administered integrated test tasks, and the relationship between task performance and test task motivation. Hundred and four learners aged between 11 and 15 years completed three computer-administered assessment tasks: a Listen-Write task, which required a summary of a listening text, and two Listen-Speak tasks, in which learners had to retell a listening text with academic and non-academic content, respectively. Participants also filled in a task-motivation questionnaire, containing items on appraisals of task difficulty, task-related emotions and anxiety, effort and subjective competence. The results indicated that the young learners held positive views on the integrated assessment tasks. Nevertheless, they found the Listen-Speak tasks significantly more difficult, more anxiety-provoking and less enjoyable than the Listen-Write task and they judged their competence to be lower than in the Listen-Write task. Task-motivational factors accounted for a low level of variation in task performance. These findings have important implications for the design and use of computer-administered integrated tasks in assessing young L2 learners.

Introduction

Young learners constitute a large part of the global population of language learners (Butler, 2017), with many instructed classroom settings focusing on English language learning. Also, in many parts of the world, an adequate level of English language proficiency may grant children opportunities for further or higher education, employment, social mobility and access to information and entertainment. The assessment of children's language abilities in instructional contexts and their results may therefore significantly affect children's life chances. This necessitates the development of valid and reliable language proficiency tests for young learners. Assessment tools are also often used to inform decisions at the level of language policy, curriculum and syllabus design, as well as for placement purposes (Wolf & Butler, 2017). Therefore, the development of high-quality language proficiency tests for young learners has recently become the focus of several international projects (for reviews see Nikolov, 2016; Wolf & Butler, 2017). These endeavours aim to design tests that take young learners' cognitive and affective developmental characteristics into consideration (Bailey, 2017), and many recent young learner language proficiency tests have been analysed in terms of their appropriacy and validity in different educational contexts (Papageorgiou & Cho, 2014; Wolf & Butler,

CONTACT Marije Michel m.c.michel@rug.nl Rijksuniversiteit Groningen, Groningen 9700 AB, Netherlands

Supplemental data for this article can be accessed here.

2017). Nonetheless, only a few studies have investigated young learners' affective reactions to these tests, and how affective factors such as motivation and test anxiety influence young learners' performance on different types of language assessment tasks (cf. Basca & Csíkos, 2016; Djigunović, 2016; Lee & Winke, 2018). While affective factors have been recognised as impacting on test performance in general (Bachman & Palmer, 2010), the examination of task-related affective factors seems particularly important in young learner contexts (Carless & Lam, 2014), because children's achievement on a test is closely linked to their interest in test tasks (Bailey, Heritage, & Butler, 2014; Hasselgreen & Caudwell, 2016) and their emotional reactions to these tasks (Jang, Vincett, Vander Boom, Lau, & Yang, 2017). In addition, in instructed classroom settings, individual instances of assessment might be considered low-stakes at face value, whereas at the same time scores might inform aspects of the teaching and learning, and thus confidence in the test results is also essential in such contexts. However, if students are not motivated to complete such low-stakes assessment tasks, the validity of the interpretation of the test scores might be compromised (Eklöf, 2010; Wise & DeMars, 2005). Therefore, the investigation of motivation seems also relevant in low-stakes test contexts.

Our research aimed to fill the above gaps and examined young English language learners' affective reactions to two types of computer-administered integrated skills test tasks in a low-stakes assessment context. In both types of tasks, which constitute part of the TOEFL® Junior™ Comprehensive test battery (Educational Testing Service – ETS), young learners in Hungary (aged between 11 and 15 years) listened to a text and then summarized what they heard in either writing or in speech. We were interested in discovering what characterizes these young language learners' test task motivation, how task motivation differs in the two types of test tasks and how task motivation relates to students' performance on the test tasks.

Review of literature

Test and task motivation

The ultimate aim of language assessment is to gain a valid and reliable measure of test-takers' second language (L2) skills and to minimize the role of additional factors that can influence test performance and which potentially contribute to measurement error. This is particularly important in the case of young learners whose working memory capacity and attentional regulation mechanisms are still undergoing development (Cho & So, 2014), and whose motivation to complete test tasks can be particularly sensitive to the assessment tasks and test administration procedures (McKay, 2006). Test motivation, that is, willingness to engage in working on test items and to invest effort and persistence in this undertaking (Baumert & Demmrich, 2001, p. 441), might have a substantial impact on test performance, and low levels of test motivation can pose a threat to the validity of the interpretation of test results.

As discussed by Baumert and Demmrich (2001), test motivation can be conceptualized as a type of achievement motivation, in other words as effort and persistence in achieving and demonstrating competence in a specific task or skill. Test motivation is a type of task motivation where students' motivational behaviour is directed towards an assessment task. Therefore, the concepts of test and test-task motivation has been developed based on theories of task motivation (e.g. Eccles et al., 1983; Eccles & Wigfield, 2002). Test motivation in general, as well as specific test task motivation is predicted by how students value the given activity – called *subjective task values*, and what their expectations are about how well they will do on the task – termed *task performance expectancy*. Test and test task performance expectancy is thereby defined as one's self-concept of domain-specific knowledge relevant for successful test and task performance (Bandura, 1997), while subjective task value is understood to be "a function of both the perceived qualities of the task and the individual's needs, goals, and self-perceptions" (Eccles et al., 1983, p. 90). Subjective task value consists of four components: attainment value, intrinsic or interest value, utility value, and cost (Eccles & Wigfield, 2002). Attainment value is related to the importance of the test and the test task for the students'

self-concept and self-expression (Eccles & Wigfield, 2002). Intrinsic or interest value is defined as "the enjoyment the individual gets from performing the activity or the subjective interest the individual has in the subject" (Eccles & Wigfield, 2002, p. 120). Utility value expresses the usefulness of a test and test task in relation to learners' current or future plans and goals (Eccles & Wigfield, 2002), and the fourth component, i.e. cost, embodies anticipated failure or potential stress and the anxiety felt during test and test task performance.

Test motivation has been traditionally measured with self-report questionnaires. One of the most frequently used instruments is the Student Opinion Scale (SOS) (Sundre & Moore, 2002), which taps into effort invested in the test (five Likert-scale items) and personal relevance of the test (five Likert-scale items). Eklöf (2006) devised her Test-Taking Motivation Questionnaire (TTMQ) based on the SOS, expanding it to a 24-item instrument with four- and five-point scales and an open-ended item. The statistical analysis of the survey responses revealed that the instrument covers two factors: performance expectancy and test-taking motivation, which includes items on effort, motivation to do well and the personal importance of the test. In a later study, Eklöf and Nyroos (2013) also added a four-item test anxiety scale to their test motivation battery. However, to our knowledge no validated instruments exist that contain an intrinsic and interest value scale, which would be needed to assess all four relevant theoretical components of test motivation. Another limitation of existing questionnaires on test motivation is that they refer to the complete test battery and do not target specific test-tasks, which constitute the focus of our study.

Research in the field of educational psychology and task-based learning has demonstrated that task characteristics can have a significant impact on task motivation. Personally relevant and novel task content can evoke situational interest, which in turn can enhance enjoyment and attentional processing and help students to sustain engagement (Hidi, Renninger, & Krapp, 1992). Poupore's study (2014), for example, which was conducted with Korean adult learners of English, demonstrated that personally relevant task content made language learning tasks interesting and enjoyable. Conversely, tasks with content that was unfamiliar and distant for the students evoked low levels of motivation. Poupore's study also highlighted that level of clarity, ease of comprehension, and concreteness of task content were important predictors of task motivation. These findings in the area of task-based learning also have high relevance for furthering the understanding of test- and test-task motivation.

Task motivation has been a relatively neglected area of research in the field of second language learning, teaching and assessment. Most existing studies only use a short task-appraisal questionnaire which assesses interest, task-related anxiety, task motivation, and perceived success in task-completion with one questionnaire item only (e.g. Kormos & Préfontaine, 2017). One of the exceptions to this is Mozgalina (2015) who examined task motivation with the help of a questionnaire that measured four components of self-determination as they relate to tasks: intrinsic motivation, identified regulation, external regulation and amotivation (Deci & Ryan, 1985). Although the questionnaire was shown to be reliable and valid, it does not directly tap into task anxiety and expanded effort. Another instrument was applied by Poupore (2014) who adopted Boekaerts' task motivation questionnaire (2002), which is one of the most widely used instruments for assessing task motivation in educational psychology. Boekaerts' questionnaire, which we also used in our study, incorporates the key theoretical components of test and task motivation outlined by Baumert and Demmrich (2001; see above): task appraisal, reported effort, emotional state, and result assessment.

The impact of test and task motivation on performance

The examination of test motivation is particularly important in low-stakes assessment contexts where test performance outcomes might be perceived to have limited direct personal relevance. In these testing situations, students' evaluation of the utility value of the test might differ, which can result in substantial variation in effort and persistence in completing test tasks (Eklöf, 2010). Consequently, the information gained in low-stakes assessment might underestimate test-takers' abilities. This can have negative

consequences for programme evaluation and, if the results are used for setting standards, for the validity of high-stakes assessment as well (Finn, 2015). Low levels of test motivation can also result in guessing, and reduce the reliability of the test task (Wise & DeMars, 2005).

Although in general educational psychology, task motivation towards learning activities has been found to be a significant predictor of task engagement and subsequent achievement (e.g. Harackiewicz, Durik, Barron, Linnenbrink-Garcia, & Tauer, 2008), the results of studies examining the role of test motivation in performance on low-stakes assessments in general education have been mixed. For example, in Wise and DeMars' (2005) review of previous research, test-takers with high motivation were found to outperform those with low test motivation by 0.5 SD. A number of studies, however, especially those that manipulated test motivation experimentally in general educational tests, showed only a relatively weak link (e.g. Liu, Bridgeman, & Adler, 2012) or no relationship between test motivation and performance (e.g. Eklöf, 2010). Therefore, further research is needed to investigate how test motivation in low-stakes second language asssessment contexts can impact on performance and thereby on the validity of test scores.

The role of task and test motivation in assessing young learners

The study of task motivation has high importance in the case of young learners as they are known to demonstrate large variability in performance across tasks at a given point in time (Alibali & Sidney, 2015). Affective factors, including task motivation, are key contributors to this variation because young learners' ability to exercise control over their emotions is still undergoing development (Butler, 2017). Previous research in the field of general education has revealed significant links between motivation, emotions and anxiety on the one hand and young learners' test performance on the other (for a review see McDonald, 2001).

The role of affective factors also seems particularly relevant in the assessment of young learners' L2 skills. Recent work in the field of L2 test development has consistently emphasized the need to take into account young learners' interest, motivation and emotions in designing assessment tasks for this learner group (for a review see Wolf & Butler, 2017). In addition, in a recent study Winke, Lee, Ahn, Choi, Cui, and Yoon (2018) demonstrated that if children are unfamiliar with the language test tasks, lack of clarity of test instructions can create construct-irrelevant variance, negative attitudes and anxiety in children. Consequently, it is important to investigate how young L2 learners' performance might be influenced by test task motivation. As test motivation was found to exert a stronger impact on more challenging tasks than on easier ones (for a review see Finn, 2015), it is particularly relevant to understand what characterizes young learners' motivation for completing cognitively demanding and complex tasks that involve the integration of various languages skills.

Affective reactions in computer-mediated testing environments

Over the past two decades, we have increasingly seen the introduction of computer-based language assessments. Previous research suggests, however, that for example speaking in computer-administered testing contexts, where tasks are performed in a closed online context, might be particularly anxiety provoking (cf. Segalowitz & Trofimovich, 2012). Unlike in face-to-face communication, test-takers do not only need to make sure that they do not speak longer than required but that they fill the time with speech and start talking at the right time. Specifically with reference to young language learners, Lee and Winke (2018) found that the absence of a real audience and lack of feedback in online speaking tasks put additional pressure on the young learners and can induce predominantly negative emotional reactions to speaking tasks.

A particular task format that is often incorporated in computer-administered L2 tests are integrated tasks, which require the employment of various language skills to complete the task (e.g. reading, listening, and writing). Nonetheless, with the exception of a recent study by Hsieh and

Wang (2019), research on integrated tasks has primarily focused on their use in adult L2-learner contexts (see e.g. Frost, Elder, & Wigglesworth, 2011; Plakans, Gebril, & Bilki, 2016). However, as Hasselgreen and Caudwell (2016) point out, strict separation of the four language skills no longer reflects language use in, for example, the internet and social media environments in which young learners interact in an L2. Furthermore, although integrated tasks have been introduced in assessment as a way to represent language use in real-life situations, even in adult L2 research the emphasis has been on cognitive and discourse features of integrated performances. Therefore, little is known about the affective dimensions of computer-administered integrated tasks in general and about the role of motivation in young learners' L2 performance on such tasks more specifically, including in low-stakes assessment contexts.

Our research aimed to fill this gap and investigated the task motivation of young language learners while completing two novel forms of computer-administered integrated assessment task types: one task type in which students summarize orally what they hear in a listening text (Listen-Speak) and another task type in which they summarize a listening text in writing (Listen-Write). We were interested in the similarities and differences between motivation towards these two integrated task-types as they differ in output modality, and due to the time-pressured nature of speaking, in perceived difficulty (Brown, Iwashita, & McNamara, 2005). Computer-administered speaking tasks might also prompt higher levels of anxiety (cf. Segalowitz & Trofimovich, 2012). Therefore, insights gained from the analysis of these two integrated skills assessment tasks can be particularly helpful for test design and the preparation of candidates for the test.

Our study asked the following research questions:

(1) What characterizes different aspects of task motivation of young learners of English in computer-administered integrated Listen-Speak and Listen-Write test tasks?
(2) Is there a difference between the motivation of young learners of English to perform computer-administered Listen-Speak versus Listen-Write test tasks?
(3) How are different aspects of task motivation in performing computer-administered Listen-Speak and Listen-Write test tasks interrelated?
(4) What is the relationship between task performance and task motivation in computer-administered Listen-Speak and Listen-Write test tasks?

Method

Participants

The participants were 104 learners of English as an additional language in two primary schools in Budapest. In these schools, children start learning English from Grade 1 and have five English language classes per week all through their primary school education (up to Grade 8). They also receive content-based instruction through the medium of English in arts, music, science, and physical education in lower primary school (Grades 2 to 4), and in history and science in upper primary school (Grades 5 to 8). Both schools are state-owned and education is free of charge.

Sixty-two of the children were from School A and 42 from School B; 43% were boys and 57% girls. The youngest participant was 11 and the oldest 15 years old, with a mean age of 12.72 (SD = 0.81). Fifty per cent of the children attended Grade 6, 43% Grade 7, and 7% Grade 8. Students' proficiency ranged from A2 to B2 level on the CEFR as measured by students' scores on the full TOEFL® Junior™ Comprehensive test-battery administered as part of a larger scale project: 31% of the children were at A2 level in terms of overall English language proficiency, 28% at B1 level, and 41% at B2 level.

Instruments

Language tasks

The participants completed three integrated language tasks for the purposes of this study. These were computer-administrated and formed part of the larger TOEFL® Junior™ Comprehensive test battery which measures all four skills needed to communicate successfully in English. The writing and speaking subtests each consist of four task types, of which we used the integrated tasks as a specific focus in this study: Academic and Non-Academic Listen-Speak tasks, and the Academic Listen-Write task.

Both the Academic and Non-Academic Listen-Speak tasks start with a brief introduction explaining the context of the listening tasks. Following the introduction, in the Non-Academic version of the task, the test-takers listen to a teacher explaining an event in regular school life (e.g. a visit to a library). The teacher might be interrupted by a short question from a pupil in the recording, but the approximate 1.5 minutes of speaking input time is primarily taken up by the teacher. In the Academic Listen-Speak task, test-takers hear a teacher giving a presentation on an academic topic (e.g. the metamorphosis of a tadpole into a frog). The presentation lasts for 2 minutes and the students can listen to it only once. In both Listen-Speak tasks, the teacher's explanations are accompanied by some pictures illustrating key terms (e.g. tadpole, froglet), and test-takers are allowed to take notes while listening. When the teacher in the recording stops speaking, test-takers receive brief instructions on how to complete the speaking task, which requires them to summarize what they heard. They then have 45 seconds to prepare their answer and one minute to record their summary while speaking into a microphone. Altogether, each task lasts about 5 minutes.

In the Listen-Write task, the introduction lasts for about 30 seconds and explains the full task. Then, test-takers listen to a teacher discussing an academic topic (e.g. reasons for differences in the amount of rainfall) while they see a picture with animations, which includes key information. The presentation lasts for about 1.5 minutes and test-takers are allowed to take notes while listening, but they can only listen to the recording once. Participants then hear the instructions for the task, which ask them to write a summary paragraph and check their responses for grammar and spelling. While typing the paragraph, the task instructions and the illustration on the computer screen remain visible on the screen, i.e. next to the text box where they have to type their answer. Test-takers have 10 minutes to write the summary. Together with the instructions, this task lasts for about 15 minutes. Thus, in total, participants spend about 25 minutes on the three tasks together.

Participants' performances were scored by ETS-trained raters based on the regular TOEFL® Junior™ Comprehensive performance descriptors for each skill.[1] The scoring rubrics take into account linguistic accuracy, the complexity and variety of language produced, the clarity and coherence with which relevant and appropriate key information from the source text is summarized. Accordingly, the integrated task performances could achieve a maximum score of 4, where the top level indicates that the test-taker has produced accurate, fluent (for speaking) and coherent (for writing) language by using simple and complex sentences to provide key information. A top score also demonstrates that the test-taker understood and accurately conveyed key ideas and sufficient supporting detail. A score of zero represents either no response or a response that is off-task.

The test had low stakes for the participants as their results did not contribute to their course grades and only group-level aggregated feedback was made available to the language teachers and the school. Although the students individually received an official TOEFL® Junior™ Comprehensive certificate of their overall language proficiency based on their test results by post to their home address, the performance outcomes on the test had no direct impact either on the students' grades or on their school achievement.

[1]ETS employs strict quality control of rating procedures and rater monitoring.

Task-motivation questionnaire

Participants' task motivation towards the two integrated task types (Listen-Write and Listen-Speak) was assessed using a questionnaire adopted from Boekaerts (2002), which is one of the most widely used instruments for assessing task motivation. This questionnaire was also selected because it incorporated the key theoretical components of test and task motivation outlined by Baumert and Demmrich (2001; see above). Boekaerts' (2002) original questionnaire taps into Task Appraisal, Reported Effort, Emotional State, and Result Assessment. These dimensions can be mapped onto the concepts of test task appraisal, test task effort, emotional dispositions towards test tasks, and judgements of competence in the test task.

In the questionnaire development phase, we first wrote the questionnaire items in English, which were then translated in the participants' L1 Hungarian by the first author of the paper whose L1 is also Hungarian. The questionnaire was set up using the software Qualtrics, which allowed us to administer it using an online interface and download the data in an electronic format.

The questionnaire consisted of two main sections that focussed on specific task-types: one section on the Listen-Write task, and the other section on the Listen-Speak task type. Both of these main sections contained the same questions. Within each main section, there were three groups of questions. The first group of questions (8 items) aimed to assess young learners' feelings after completing the task on a four-point semantic differential scale. We adopted six items from Boekaerts' (2002) nine-item scale. The scale contained adjectives we judged young learners would be able to understand and distinguish from other types of feelings easily, and for which translations into Hungarian were relatively straightforward (e.g. worried, confident). We also added two more items to Boekaerts' scale that expressed emotional states young learners would easily be able to relate to (happy, bored).

The second group of task-motivation questions consisted of five different four-point Likert scale items which were also adopted from Boekaerts' (2002) questionnaire (originally seven items). These questions were presented as multiple-choice questions (with the options representing different Likert-scale points). The items focused on the dimensions of Reported Effort, Task Appraisal, Emotional State and Result Assessment and asked learners about the effort they invested in task completion, perceived task difficulty, their self-evaluation of success and their enjoyment of the task.

The third group of questions were seven statements (also adapted from Boekaerts, 2002; originally eight items) for which the participants needed to indicate their level of agreement on a four-point Likert-scale. These items represented the dimensions of Reported Effort, Task Appraisal, Emotional State, and Subjective Competence.

The questionnaire also included additional items for the Listen-Speak task specifically where we asked whether the participants felt there was a difference between the Academic and the Non-Academic Listen-Speak tasks, and if so, in what ways these two Listen-Speak tasks differed. The learners were thereby prompted to consider issues related to their emotional state, task difficulty, invested effort, topic familiarity and result assessment. A copy of the questionnaire can be found in the supplementary online materials.

Procedures

An initial draft of the task-motivation questionnaire and the TOEFL® Junior™ Comprehensive test were piloted with 14 participants with similar background characteristics and in the same context as in the main study. In the pilot version of the questionnaire, learners were asked about their views on the Academic Listen-Speak task and the Non-Academic Listen-Speak task separately. In other words, the task-motivation questions were repeated for each Listen-Speak task. However, the pilot results indicated that the learners were not able to distinguish or differentiate their answers between the Academic and Non-Academic tasks, providing the same answers for both. Thus, for the main study it was decided to collapse the task-motivation questions on the two Listen-Speak tasks and elicit young learners' views on both of these together. To compensate for this, in the main study, we added

an additional open-ended question on what differences the learners perceived in the two versions (Academic/Non-Academic). This also had the advantage of a shorter questionnaire and yielded some qualitative data. The open-ended answers in the main study showed that the students could distinguish the two tasks quite successfully. The analysis of the descriptive statistics of the questionnaire as well as the informal feedback of the pilot participants suggested that no other major modifications were needed to the research instruments. The pilot study also indicated that the test was appropriate for the target group of learners in terms of language proficiency level, structure, timing, and computer literacy skills.

Prior to administration of the instruments, ethical approval for the research was granted by the relevant ethics review committee at the researchers' institution, and consent was sought from parents as well as the young learners themselves. They were provided with an information sheet in Hungarian and the study was also explained orally by a research assistant. In addition, the participants' classroom teachers conducted a familiarization session with the learners before the test, using publicly available sample materials and the TOEFL® Junior™ Comprehensive test handbook.

On the data collection day, the participants first completed the TOEFL® Junior™ Comprehensive test, which was 2 hours and 14 minutes long. Next, the participants filled out the task-motivation questionnaire[2] and a short bio-data questionnaire which together took about 10 minutes to complete. Throughout the experiment, the participants were given a number of breaks (in line with the test regulations) and provided with light refreshments.

Analyses

In order to answer our research questions, we first carried out a principal component analysis to examine the structure of the hypothesized task-motivation scales.[3] The principal component analysis of the questionnaire responses revealed that the items that were originally designed to tap into the construct of emotional state needed to be divided into the separate constructs of task-related emotions and a specific sub-type of emotions: task-related anxiety. The analyses also showed that two items referring to how relieved the participants were and how satisfied they felt after they completed the tasks had to be omitted due to low commonality values. The final factor solution for task-related emotion and anxiety constructs met the required statistical criteria. For the items referring to the Listen-Speak tasks, the KMO value was .701, and Barlett's Test of Sphericity reached statistical significance ($p < .001$). The two components with eigenvalues exceeding 1 accounted for 38.81% (task-related emotions) and 20.62% (task-related anxiety) of the variance, respectively (the total variance explained was 65.41%) (see supplementary online material for the factor loadings for these two scales). With regard to the items referring to the Listen-Write task, the KMO value was .742, which is above the recommended value of .50 (Pett, Lackey, & Sullivan, 2003) and Barlett's Test of Sphericity reached statistical significance ($p < .001$). The principal components analysis revealed the presence of two components with eigenvalues exceeding 1, explaining 36.78% (task-related emotions) and 22.97% (task-related anxiety) of the variance respectively (the total variance explained was 64.72%). We also checked the reliability of these two scales. They all had acceptable Cronbach alpha values above .700 (task-related anxiety: [nervous, worried, confident] Cronbach alpha: Listen-Speak: .706, Listen-Write: .701; task-related emotions [fed-up, bored, happy, enjoyable, in the mood] Cronbach alpha: Listen-Speak .773, Listen-Write: .795). Average scores were computed for the task-related emotions and anxiety for each of the tasks separately.

[2]Due to the timed nature of the computer-based assessment and the technical requirements of test administration, it was not possible to administer the task-motivation questionnaire immediately after students performed each task.

[3]The sample size for conducting principal component analysis was deemed sufficient based on Mundfrom, Shaw, and Ke (2005), who recommend a minimum of 75 participants for two factors with nine variables for good "agreement with the population structure from which the sample was taken" (p. 161) and 150 for excellent agreement.

The three items originally hypothesized to constitute the reported effort scale did not form one meaningful factor. One item ('I did my best in this task') was not significantly related to the items enquiring into effort expended on the task and the attention paid to completion of the task. The two items on effort and attention, however, were significantly and moderately correlated in both the Listen-Speak ($r = .505$, $p < .001$) and Listen-Write tasks ($r = .426$, $p < .001$). Therefore, we created an average score for reported effort based on learners' answers to the two questions on effort and attention only.

The two items tapping into task appraisals regarding the difficulty of the two task types were also significantly correlated in the Listen-Speak ($r = .669$, $p < .001$) and Listen-Write tasks ($r = .526$, $p < .001$). Therefore, these two items were averaged to create a task-appraisal score referring to each task.

Following the principal component analysis, we examined the descriptive statistics for the task-appraisal, task-related emotion, task-related anxiety and task effort scales and of the single item that tapped into subjective competence (see Research Question 1). For all these scales and items mean scores close to 1 represented 'low' and close to 4 'high' values (where logically needed, items were reverse-coded). Data for many of the investigated scales and items were skewed and showed a somewhat restricted range of distribution. Therefore, in all further inferential statistical analyses, non-parametric tests were used.

In order to find out if statistically significant differences exist between the two task types in terms of task-motivational factors (see Research Question 2), we applied Wilcoxon Signed Ranks tests.

With regard to the inter-relationship of task-motivational factors (see Research Question 3), we first ran Spearman rank-order correlations to examine whether task-motivation factors were related in the two tasks. Following Cohen (1988), correlation coefficients between .10-.29 were evaluated as small, between .30-.49 as moderate and above .50 as strong. In order to correct for Type I error arising from multiple comparisons, the significance level of correlational analyses was set at $p < .001$.

To investigate the relationship between task performance and task-motivational factors (see Research Question 4), we first examined the box-plots of task performance scores. They showed that there were outliers in all three tasks, and these outliers scored zero (n = 1) and one point (n = 1) in the Non-Academic Listen-Speak and one point (n = 5) in the Academic Listen-Speak tasks, and zero (n = 2) and one point (n = 1) in the Listen-Write task. We excluded the participants with a score of zero (n = 3) from the correlational analyses which were used to assess the link between task-motivational factors and task scores.

Results

Task motivation in listen-speak and listen-write test tasks

In our first research question, we were interested in discovering what characterizes the task motivation of young learners of English in computer-administered integrated Listen-Speak and Listen-Write test tasks. In order to be able to evaluate motivation towards these two test tasks in relation to students' performance on the tasks, we also examined whether students' scores differed on the Listen-Speak and Listen-Write test tasks. The repeated measures ANOVA analysis showed that learners' scores on the Academic and Non-Academic Listen-Speak and Listen-Write tasks was not significantly different $F(2, 102) = 2.14$, $p = .122$.

As regards task appraisals, the learners judged the test tasks of moderate difficulty with a mean of 2.63 for the Listen-Speak tasks and 2.85 for the Listen-Write task. The mean values also indicated high levels of reported effort and positive emotions towards both types of test tasks (see Table 1). Their reported level of anxiety was in the moderate range, and they judged their own competence in completing these test tasks relatively highly.

As can be seen in Table 1, except for task effort, all the task-motivational factors were significantly different from each other (see Research Question 2). The learners appraised the Listen-Speak tasks as

Table 1. Descriptive statistics for the task-motivation variables (N = 104).

	Mean out of 4	Median	*SD*	Skewness	Kurtosis	*Z*	*p*
Task-appraisal Listen-Speak	2.63	3.00	.69	−.16	−.05	−3.39	.001
Task-appraisal Listen-Write	2.85	3.00	.46	−.57	.58		
Task-effort Listen-Speak	3.19	3.00	.58	−.17	−.47	−.85	.395
Task-effort Listen-Write	3.16	3.00	.53	−.59	1.53		
Task-anxiety Listen-Speak	2.36	2.00	.80	.22	−.57	−5.02	<.001
Task-anxiety Listen-Write	1.95	2.00	.74	.66	−.07		
Task-emotion Listen-Speak	3.08	3.00	.60	−.56	.40	2.24	.025
Task-emotion Listen-Write	3.16	3.00	.59	−1.19	2.06		
Subjective competence Listen-Speak	2.91	3.00	.76	−.38	−.07	−3.84	<.001
Subjective competence Listen-Write	3.19	3.00	.56	.04	−.11		

significantly more difficult and perceived them to be more anxiety provoking than the Listen-Write task. Task-related emotions were also more positive for the Listen-Write task than the Listen-Speak tasks. They judged their own abilities to complete the Listen-Write task higher than for the Listen-Speak tasks.

As regards the inter-relationship of task-motivational factors (see Research Question 3), the analyses reported in Table 2 show that learners' task motivation in the two assessment task types was strongly related. Those who judged the Listen-Speak tasks difficult and anxiety provoking also found the Listen-Write task challenging and stressful. Learners' evaluation of task-related emotions and their task efforts were also strongly associated on the two test tasks.

Further Spearman rank-order correlations were run to examine the inter-relationships among the task-motivational factors within each assessment task (see Table 3). In the Listen-Speak tasks, task appraisal correlated moderately with task-anxiety and task-emotions, and more strongly with subjective competence. Moderate associations were also found between task-emotion and subjective competence, and task-emotion and task-effort. The strength of association between task anxiety and subjective competence was on the borderline between weak and moderate.

The analysis of the Listen-Write task revealed a similar pattern of correlations (see Table 3), but the strength of the association was weaker between task-appraisals on the one hand and task-related

Table 2. Relationship between task-motivation factors in the listen-write and listen-speak tasks.

Variable pairs	Spearman's rho	*p*
Task-appraisal Listen-Speak & Task-appraisal Listen-Write	.527	<.001
Task-effort Listen-Speak & Task-effort Listen-Write	.590	<.001
Task-anxiety Listen-Speak & Task-anxiety Listen-Write	.519	<.001
Task-emotion Listen-Speak & Task-emotion Listen-Write	.719	<.001
Subj. competence Listen-Speak & Subj. competence Listen-Write	.489	<.001

Table 3. Spearman rank-order correlations between task-motivation factors in the listen-speak and listen-write tasks.

		Task-effort	Task-anxiety	Task-emotion	Subj. competence
Listen-Speak	Task-appraisal	−.045	−.452[b]	.408[b]	.530[b]
	Task-effort		.095	.334[b]	−.032
	Task-anxiety			−.304[b]	−.298[b]
	Task-emotion				.417[b]
Listen-Write	Task-appraisal	.003	−.246[a]	.202[a]	.335[b]
	Task-effort		−.038	.489[b]	.101
	Task-anxiety			−.163	−.272[b]
	Task-emotion				.365[b]

[a]$p < .05$
[b]$p < .001$

anxiety, emotions and subjective competence on the other. Subjective competence was also weakly associated with task-related anxiety. Moderately strong correlations emerged between task-related emotions, and task effort and subjective competence.

Relationship between task motivation and test task performance

In order to answer our fourth research question, we first examined the descriptive statistics for the Academic and Non-Academic Listen-Speak tasks and the Listen-Write task. The mean values for each of the three tasks were around a score of 3 out of a maximum of 4 (see Table 4). This indicates that, overall, the young learners performed quite well on the three assessment tasks. The correlational analyses of the task scores and task-motivation factors revealed that subjective evaluations of competence were associated with the Listen-Write and Non-Academic Listen-Speak task scores, and that task-appraisals correlated with Non-Academic Listen-Speak task scores (see Table 5). However, all these correlations were weak and none of them were significant when the *p* values were adjusted for multiple comparisons.[4]

Table 4. Descriptive statistics for the task performance variables (N = 104).

	Mean	*SD*	Skewness	Kurtosis
Listen-Speak Non-Academic	3.17	.75	−.98	2.11
Listen-Speak Academic	2.99	.81	−.52	−.15
Listen-Write	3.10	.83	−1.12	2.24

Table 5. Spearman rank-order correlations between task-motivation factors and task performance.

	Non-Academic Listen-Speak	Academic Listen-Speak	Listen-Write
Task-appraisal	.209[a]	.138	.133
Task-effort	−.100	.136	.157
Task-anxiety	−.123	−.081	−.068
Task-emotion	−.110	.087	.058
Subjective competence	.208[a]	.111	.217[a]

[a] $p < .05$

Discussion

In our research, we first examined what characterizes the task motivation of young learners when completing computer-administered integrated assessment tasks. We identified five distinct, but inter-related, constructs of task motivation which correspond to key components of task and test motivation (Baumert & Demmrich, 2001). Task-appraisals represent the attainment and utility value of test tasks, task-related anxiety embodies the perceptions of cost, task-related emotions express the intrinsic value of tasks, and subjective competence describes learners' self-efficacy beliefs, namely to task attainment value. Task effort yields information about persistence with the test tasks. Overall, the descriptive statistics for the task-motivation scales indicated that our participants had favourable motivational dispositions towards both integrated assessment task-types. The learners considered the tasks to be of moderate difficulty and within their competence. The participants enjoyed performing the integrated skills tasks, and reported that they expended sufficient effort on them. According to their self-report, participants displayed only a moderate level of anxiety.

One of the possible explanations for the learners' positive evaluations of the computer-administered integrative assessment tasks, and their own abilities and feelings, might be the careful

[4]Using a MANOVA analysis, we found no association between CEFR level and task motivational factors F (3, 100) = 1.01, p = .391, Wilks lambda = .729.

design of the tasks that took into account young learners' interests and cognitive abilities. So et al. (2017) explain that in the development process of the TOEFL® Junior™ Comprehensive test, tasks were chosen so that they are developmentally appropriate and interesting in terms of both content and task structure and that they reflect target language use in English-medium instruction contexts. Indeed, the school-related topics and scenarios in the Listen-Speak and Listen-Write tasks reflect Hasselgreen and Caudwell's (2016, p. 1) observation that "[s]chool is the most normal arena for language learning, and frequently also for language use" for young L2 learners. Hidi and Harackiewicz (2000) furthermore highlight that tasks that have a moderate level of difficulty and complexity, and which are novel and personally meaningful, arouse situational interest and enhance engagement (cf. Poupore in the field of L2 task-based research). Similarly, Malloy (2015) emphasizes the positive role of more authentic and relevant tasks – which integrated test tasks have been argued to be. The fact that the tasks were computer-administered rather than presented in a traditional paper-based format might also have contributed to positive task motivation.

The test tasks in this study were accompanied by short clear instructions, which is another feature that Malloy (2015) and Winke et al. (2018) have put forward as vital to helping young learners understand task relevance and helping to control for stress and anxiety. Winke et al. (2018) also highlighted the importance of test familiarity in fostering positive affective reactions to language tests. In our research, all participants were familiarized with the types of tasks prior to the study by their teachers in a session that involved completing a set of sample test tasks. This ensured that the task types were known to the learners and that they had an appropriate understanding of what they needed to do during the assessment process. Indeed, test familiarity is another important factor that is known to contribute positively to test motivation (Baumert & Demmrich, 2001). Furthermore, Winke et al. (2018) pointed out that test wiseness can also threaten the validity of the assessment, especially in cases when there is variation among test-takers in familiarity with the tasks. Thus, the opportunity to prepare for a test is a key component of test fairness (American Educational Research Association [AERA], 2014), therefore it is essential to ensure that – regardless of the stakes of a test – examinees are provided with equal chances to familiarize themselves with the test tasks and procedures (see also Winke et al., 2018 for a similar argument).

When comparing the two task types, the results indicate that our participants judged the Listen-Speak tasks to be significantly more difficult, more anxiety-provoking and less enjoyable than the Listen-Write task, and that they perceived their competence to be lower in the former than in the integrated-writing task (cf. Research Question 2). However, our analysis showed that learners' *actual* scores on the Academic and Non-Academic Listen-Speak and Listen-Write tasks did not differ. A possible reason for the difference in the task-motivational components might be related to the fact that L2 learners may feel somewhat more anxious and less self-confident when speaking because it is a time-pressured activity (cf. Segalowitz & Trofimovich, 2012). The less favourable motivational dispositions might also be due to the computer-mediated mode of speaking test delivery. As in Lee and Winke's (2018) study, our participants found it somewhat unusual and stressful to speak to a computer. This was also expressed in one of the participants' comments: "It was strange to talk to a machine and I felt I did not do so well." Furthermore, some qualitative comments on the Listen-Speak task in the present study showed that participants felt that the 45 second-preparation time was not enough for them to plan what they wanted to say. In contrast, writing is less time-constrained and there are opportunities for revision. The relatively high level of perceived difficulty of the Listen-Speak task is also in line with Brown et al. (2005). Looking into the TOEFL iBT, they found that this type of task was perceived – in their case by raters – as the most challenging speaking task type for adult L2 learners. Speaking tests are less frequently administered than written tests in the Hungarian school context and this lack of experience might also have been a cause of increased test task anxiety. Although the input phase of the Listen-Speak and Listen-Write tasks is very similar in structure, the time pressure and the output mode differences can create test-task anxiety and make Listen-Speak tasks less enjoyable for young learners. Therefore, in test preparation and classroom language teaching activities more generally, children could be asked to record their summaries of audio or

audio-visual input digitally. Such podcast-type audio-recordings offer great opportunities for speaking practice, and help familiarize L2 learners with computer-administered speaking assessment tasks.

Our third research question enquired into the inter-relationship of various components of task motivation between the two assessment task types. The correlational analyses of the different motivational scales in the two tasks revealed that task-appraisals, anxiety, emotions and subjective perceptions of competence were significantly related in the two assessment tasks. In the case of task appraisals, anxiety and subjective competence, the strength of the relationship was moderate, which suggests that there is some degree of commonality in young learners' motivation towards these two types of test tasks. In the case of task-related emotions the correlation was strong, which is indicative of a larger overlap across test tasks. The shared variation in task-motivational factors may be due to the relative similarity of the processing demands of the integrated task types, which both required that young learners listen to a text, take notes and summarize information.

As mentioned earlier, a correlational analysis of the task-motivational factors also suggests that task-effort, anxiety, emotions and subjective perceptions of competence are distinct constructs and are either weakly or moderately related to each other. This supports Finney, Mathers, and Myers (2016) argument that test motivation, even in low-stakes contexts, is not a unidimensional construct. The strength of association among the task-motivational factors was somewhat higher in the Listen-Speak tasks than in the Listen-Write task. The pattern of correlations, however, was similar in both tasks, with the exception of task anxiety which only correlated significantly with task emotions in the Listen-Speak tasks and not in the Listen-Write task. The somewhat stronger link between task anxiety and emotions in the speaking task might be due to the generally more anxiety-provoking nature of speaking when compared to writing (e.g. Pae, 2013), the demands of the closed-context online speaking task (cf. Lee & Winke, 2018), and the larger range of scores on these in our sample. The strongest links in the Listen-Write task were found between task emotions on the one hand, and effort and subjective perceptions of competence on the other, which highlights the important role of young learners' emotions in task-specific and general academic effort (Valiente, Swanson, & Eisenberg, 2012).

Task-related anxiety was found to be negatively related to task appraisals and subjective competence in both types of assessment tasks. In other words, young learners who felt anxious about the integrated assessment tasks judged them to be more difficult and had more negative perceptions about their ability to complete them successfully. These findings are in line with a number of earlier studies conducted with older L2 learners that have shown that self-confidence plays an important role in writing anxiety (e.g. Woodrow, 2011) and oral communication (e.g. MacIntyre, Noels, & Clément, 1997).

Participants' subjective perceptions of their competence were also significantly correlated with task appraisals in the written and spoken version of the tasks. Judgements of competence correspond to the construct of task-related self-efficacy beliefs. Bandura (1997) argues that an important source of information of children's self-efficacy beliefs is how well they have done previously in similar tasks or task types. Previous experience of success can lead to the development of a stronger sense of self-confidence when performing similar tasks in the future. Therefore, the perceived difficulty of tasks is importantly associated with subjective perceptions of competence, as our study also shows.

Our final research question sought to uncover the relationship between task-motivational factors and test task performance. The analyses revealed that task motivation was mostly unrelated to young learners' scores in the computer-administered integrated test tasks. The weak relationships between performance level and other task-motivational variables were somewhat unexpected, given that previous research has shown the important role of task-effort, anxiety and emotions in young learners' task performance (for a review see McDonald, 2001). This finding is also contrary to the results of studies that have found an association between test motivation and test outcomes in low-stakes educational assessment contexts (for a review see Wise & DeMars, 2005). However, the fact that none of the task-motivation variables shared more than 5% of variance with the task performance scores suggests that task-motivational factors did not create substantial construct-irrelevant variance in our study. We would argue that a potential reason for this finding might be related to the

design and content of the test and the conditions in which the tasks were administered in our research. As discussed earlier, the TOEFL® Junior™ Comprehensive test battery has been developed taking young learners' cognitive characteristics, their language use domain and interests into account (So et al., 2017), and the participants took the test in the computer labs of their school, in a familiar and comfortable setting. These factors might, at least partly, explain our participants' positive task and test motivation. In our research the young learners had been made familiar with the task types and had practised them prior to data collection, which might have contributed to the relatively low level of test task anxiety. Also, although the test was low-stakes and the scores were only reported to the participants and their parents, the learners knew they would receive an official certificate, which may have encouraged them to exert high task effort. Therefore, the effect of task-motivational factors on test scores might have diminished in this low-stakes assessment context.

Conclusion

Our study concentrated on the comparatively under-researched area of young L2 learner assessment and was novel in its focus on computer-administered integrated tasks and on the nature and effect of task motivation. The results indicate that the young learners in our study held positive views on the integrated assessment tasks they completed (two Listen-Speak and one Listen-Write task from the computer-administered TOEFL® Junior™ Comprehensive test). Nevertheless, they were somewhat more favourable towards the Listen-Write task, whereas they found the Listen-Speak tasks comparatively more difficult, more anxiety-provoking and less enjoyable, and they perceived their competence to be lower in them. These different tendencies in test task motivation, however, were not reflected in the learners' actual scores on the different integrated tasks, with similar performance results on both task types.

Our study also explored the inter-relationship of various components of test task motivation between the two integrated task types. Our findings suggest that there is some degree of commonality in young learners' motivation towards these two types of assessment tasks in terms of the dimensions of task appraisals, anxiety, emotions and subjective competence. Despite earlier suggestions in the literature of an association between motivation and test performance, we did not find a substantial link between task motivation and test-task performance on the computer-administered integrated tasks investigated in our study. As discussed above and further commented on below, this finding might be due to features of the tasks and participant population of our study.

A key implication of our study and its findings for young L2 learner pedagogy and assessment is the positive reception of computer-administered integrated tasks, provided the tasks are well designed. This requires that the test tasks contain clear, simple and straightforward instructions, and that the scenarios, contexts, and content of the tasks closely relate to the learners' world experience (e.g. school setting, daily school activities and plausible classroom content). Furthermore, assessment tasks should be administered in an environment that is familiar and comfortable for young learners, and hands-on practice should be provided *prior* to the assessment event to reduce test anxiety (see also Lee & Winke, 2018). These test design and administration conditions might then ensure that variations in test motivation do not affect the validity of the interpretation of test scores.

We should point out, however, that our research was not without its limitations. One of these is that although our study involved a relatively large number of participants, an even bigger sample size would be needed to ensure more statistical power. Another limitation concerns the use of a self-reported survey instrument. Future research should consider adding a response-time effort measure that yields better insights into actual behaviour as a function of time (Wise & Kong, 2005). Also, due to technological restrictions, it was impossible to counter-balance the order in which students carried out the Listen-Speak and Listen-Write tasks. Therefore, the order of tasks might have influenced students' task motivation and performance. At the same time, we want to point out that we followed the task order as used in operational testing. Finally, it also needs to be considered that our research

was conducted in the context of English as an additional language, whereby the young learners had been learning English from the start of primary school and approximately one third of the subjects (sciences, art, music, and physical education) were taught through the medium of English. This was justified as this is the target population of the TOEFL® Junior™ Comprehensive test (see "Who uses the *TOEFL Junior* tests?" at: https://www.ets.toefl_junior/faq/). It may have meant, however, that our participants came from supportive family backgrounds with positive attitudes towards learning English, since their parents/legal guardians had enrolled them in this type of immersion school. The latter would also help explain the relatively limited range and variation in our data set, which also affect the scope of detecting significant correlations between variables. It is unknown, therefore, to what extent our findings can be generalized to a more restricted English as a Foreign Language context in which English is taught only as a subject for a limited number of hours at a later starting age. Future research would thus need to explore the nature and effect of task motivation on young learner L2 testing for such instructional contexts.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Educational Testing Service, USA undera Committee of Examiners and TOEFL research grant. ETS does not discount or endorse the methodology, results, implications or opinions presented by the researcher(s).

References

AERA/APA/NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

Alibali, M. W., & Sidney, P. G. (2015). Variability in the natural number bias: Who, when, how, and why. *Learning and Instruction*, *37*(1), 56–61. doi:10.1016/j.learninstruc.2015.01.003

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford, UK: Oxford University Press.

Bailey, A. L., Heritage, M., & Butler, F. A. (2014). Developmental considerations and curricular contexts in the assessment of young language learners. In A. Kunnan (Ed.), *The companion to language assessment* (pp. 421–439). New York, NY: John Wiley.

Bailey, A. L. (2017). Theoretical and developmental issues to consider in the assessment of young learners' English language proficiency. In M.K. Wolf, & Y.G. Butler (Eds.), *English language proficiency assessments for young learners* (pp. 25-40). New York: Routledge.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: W.H. Freeman and Company.

Basca, E., & Csíkos, C. (2016). The role of individual differences in the development of listening comprehension in the early stages of language learning. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 263–289). New York, NY: Springer.

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, *16*(3), 441–462. doi:10.1007/BF03173192

Boekaerts, M. (2002). The on-line motivation questionnaire: A self-report instrument to assess students' context sensitivity. In P. R. Pintrich & M. L. Maehr (Eds.), *Advances in motivation and achievement. New directions in measures and methods* (pp. 77–120). Amsterdam, Netherlands: JAI.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-academic purposes speaking tasks* (ETS Research Report, RR-05-05). Princeton, NJ: Educational Testing Service.

Butler, Y. G. (2017). The role of affect in intraindividual variability in task performance for young learners. *TESOL Quarterly*, *51*(3), 728–737. doi:10.1002/tesq.385

Carless, D., & Lam, R. (2014). The examined life: Perspectives of lower primary school students in Hong Kong. *Education 3-13: International Journal of Primary, Elementary and Early Years Education*, *42*(3), 313–329. doi:10.1080/03004279.2012.689

Cho, Y., & So, Y. (2014). Construct-irrelevant factors influencing young EFL learners' perceptions of test task difficulty (TOEFL research memorandum ETS RM 14–04). Princeton, NJ: Educational Testing Service.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behaviour*. London, UK: Plenum Press.
Djigunović, J. M. (2016). Individual learner differences and young learners' performance on L2 speaking tests. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 243–261). New York, NY: Springer.
Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motivation* (pp. 75–146). San Francisco, CA: W. H. Freeman.
Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109–132. doi:10.1146/annurev.psych.53.100901.135153
Eklöf, H. (2006). Development and validation of scores from an instrument measuring student test-taking motivation. *Educational and Psychological Measurement*, *66*(5), 643–656. doi:10.1177/0013164405278574
Eklöf, H., & Nyroos, M. (2013). Pupil perceptions of national tests in science: perceived importance, invested effort, and test anxiety. *European Journal of Psychology of Education*, *28*(2), 497–510.
Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice*, *17*(4), 345–356. doi:10.1080/0969594X.2010.516569
Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series*, *2015*(2), 1–17. doi:10.1002/ets2.12067
Finney, S. J., Mathers, C. E., & Myers, A. J. (2016). Investigating the dimensionality of examinee motivation across instruction conditions in low-stakes testing contexts. *Research and Practice in Assessment*, *11*, 5–17.
Frost, K., Elder, C., & Wigglesworth, G. (2011). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test-takers' oral performances. *Language Testing*, *29*(3), 1–25. doi:10.1177/0265532211424479
Harackiewicz, J. M., Durik, A. M., Barron, K. E., Linnenbrink-Garcia, L., & Tauer, J. M. (2008). The role of achievement goals in the development of interest: Reciprocal relations between achievement goals, interest, and performance. *Journal of Educational Psychology*, *100*(1), 105–122. doi:10.1037/0022-0663.100.1.105
Hasselgreen, A., & Caudwell, G. (2016). *Assessing the language of young learners*. Bristol, UK: Equinox.
Hidi, S., Renninger, K. A., & Krapp, A. (1992). The present state of interest research. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 433–446). Hillsdale, NJ: Erlbaum.
Hidi, S., & Harackiewicz, J. M. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research*, *70*(2), 151–179. doi:10.3102/00346543070002151
Hsieh, C.-N., & Wang, Y. (2019). Speaking proficiency of young language students: A discourse-analytic study. *Language Testing*, *36*(1), 27–50. doi:10.1177/0265532217734240
Jang, E. E., Vincett, M., Vander Boom, E., Lau, C., & Yang, Y. B. (2017). Considering young learners' characteristics in developing a diagnostic assessment intervention. In M. K. Wolf & Y. G. Butler (Eds.), *English language proficiency assessments for young learners* (pp. 193–213). New York, NY: Routledge.
Kormos, J., & Préfontaine, Y. (2017). Affective factors influencing fluent performance: French learners' appraisals of second language speech tasks. *Language Teaching Research*, *21*(6), 699-716. doi:10.1177/1362168816683562
Lee, S., & Winke, P. (2018). Young learners' response processes when taking computerized tasks for speaking assessment. *Language Testing*, *35*(2), 239–269. doi:10.1177/0265532217704009
Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher*, *41*(9), 352–362. doi:10.3102/0013189X12459679
MacIntyre, P. D., Noels, K. A., & Clément, R. (1997). Biases in self-ratings of second language proficiency: The role of language anxiety. *Language Learning*, *47*(2), 265–287. doi:10.1111/0023-8333.8199700
Malloy, A. (2015). Seven essential considerations for assessing young learners. *Modern English Teacher*, *24*(1), 20–23.
McDonald, A. S. (2001). The prevalence and effects of test anxiety in school children. *Educational Psychology*, *21*(1), 89–101. doi:10.1080/01443410020019867
McKay, P. (2006). Assessing young language learners. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511733093
Mozgalina, A. (2015). More or less choice? The influence of choice on task motivation and task engagement. *System*, *49*(1), 120–132. doi:10.1016/j.system.2015.01.004
Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, *5*, 159–168. doi:10.1207/s15327574ijt0502_4
Nikolov, M. (Ed.). (2016). *Assessing young learners of English: Global and local perspectives*. Berlin, Germany: Springer.
Pae, T. I. (2013). Skill-based L2 anxieties revisited: Their intra-relations and the inter-relations with general foreign language anxiety. *Applied Linguistics*, *34*(2), 232–252. doi:10.1093/applin/ams041
Papageorgiou, S., & Cho, Y. (2014). An investigation of the use of TOEFL® Junior™ Standard scores for ESL placement decisions in secondary education. *Language Testing*, *31*(2), 223–239. doi:10.1177/0265532213499750
Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Thousand Oaks, CA: Sage.

Plakans, L., Gebril, A., & Bilki, Z. (2016). Shaping a score: Complexity, accuracy, and fluency in integrated writing performances. *Language Testing*. Advance online publication. doi:10.1177/0265532216669537

Poupore, G. (2014). The influence of content on adult L2 Learners' task motivation: An interest theory perspective. *The Canadian Journal of Applied Linguistics*, *17*(2), 69–86.

Segalowitz, N., & Trofimovich, P. (2012). Second language processing. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 179–192). New York, NY: Routledge.

So, Y., Wolf, M. K., Hauck, M. C., Mollaun, P., Rybinski, P., Tumposky, D., & Wang, L. (2017). TOEFL Junior© design framework. In M. K. Wolf & Y. G. Butler (Eds.), *English language proficiency assessments for young learners* (pp. 59–78). New York, NY: Routledge.

Sundre, D. L., & Moore, D. L. (2002). Assessment measures: The student opinion scale—A measure of examinee motivation. *Assessment Update*, *14*(1), 8–9.

Valiente, C., Swanson, J., & Eisenberg, N. (2012). Linking students' emotions and academic achievement: When and why emotions matter. *Child Development Perspectives*, *6*(2), 129–135. doi:10.1111/j.1750-8606.2011.00192.x

Winke, P., Lee, S., Ahn, J. I., Choi, I., Cui, Y., & Yoon, H. J. (2018). The cognitive validity of child English language tests: What young language learners and their native-speaking peers can reveal. *TESOL Quarterly*, *52*(2), 274–303. doi:10.1002/tesq.396

Winke, P., Lee, S., Ahn, J. I., Choi, I., Cui, Y., & Yoon, H-J. (2018). The cognitive validity of child English language tests: What younglanguage learners and their native#speaking peers can reveal. *TESOL Quarterly*, *52*(2), 274-303. doi:10.1002/tesq.2018.52.issue-2

Wise, S. L., & DeMars, C. E. (2005). Low examinee motivation in low-stakes assessment: Problems and potential solutions. *Educationale first wrote the questionnaire items in Assessment*, *10*(1), 1–17. doi:10.1207/s15326977ea1001_1

Wise, S. L., & Kong, L. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183. doi:10.1080/0969594X.2010.516569

Wolf, M. K., & Butler, Y. G. (2017). An overview of English language proficiency assessments for young learners. In M. K. Wolf & Y. G. Butler (Eds.), *English language proficiency assessments for young learners* (pp. 3–22). New York, NY: Routledge.

Woodrow, L. (2011). College English writing affect: Self-efficacy and anxiety. *System*, *39*(4), 510–522. doi:10.1016/j.system.2011.10.017